横浜医学，71，577–580（2020）

総　　説（2019年度横浜市立大学医学会賞受賞論文）

大規模医療データ活用の課題に対処する臨床研究

後　藤　　温

横浜市立大学　医学群／大学院　データサイエンス研究科ヘルスデータサイエンス専攻

要　旨：近年，健診や診療データのほか，ゲノム情報やメタボローム情報をはじめとするオミックスデータが，大規模のデータベースとして蓄積されるようになった．健康医療分野において，このような大規模医療データを活用することで，様々な課題を解決することが期待されている．しかし現時点では，このような大規模医療データの分析には，多くの課題が存在する．本稿では，（1）質の高いデータを収集することによる健康医療データベースの構築，（2）正確なアノテーション情報の蓄積，（3）未観察・未測定交絡への対処，について概説し，膨大な健康医療データを活用する際の課題を克服する方法を紹介したい．

Key words:　大規模医療データ（Big Data in Healthcare），データサイエンス（data science），疫学（epidemiology），臨床研究（clinical research）

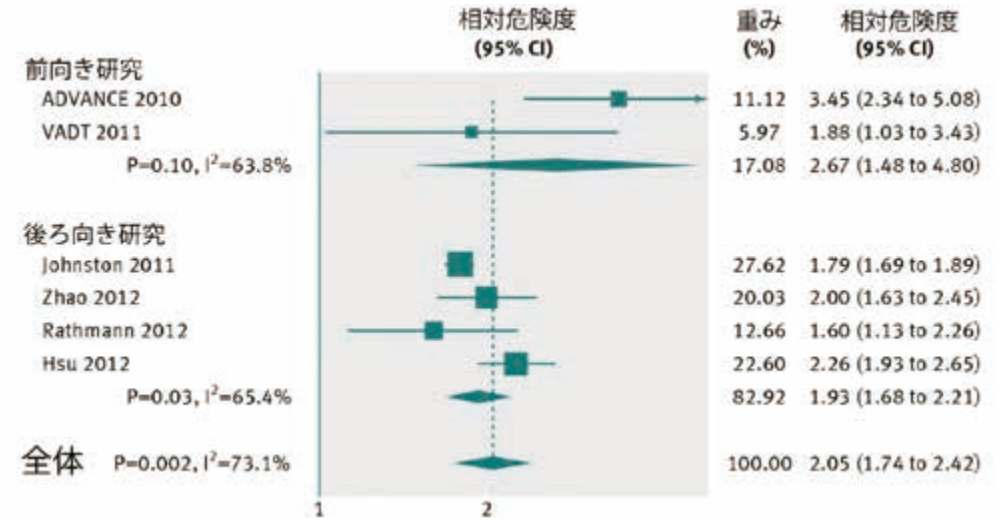
はじめに

近年，健診データ，レセプトデータ，診断群分類別包括評価（Diagnosis Procedure Combination: DPC）データやオーダリング，検査結果等を含む診療データのほか，ゲノム情報やメタボローム情報をはじめとするオミックスデータが，大規模のデータベースとして蓄積されるようになった．健康医療分野において，このような大規模医療データを活用できるようになると，そのメリットは計り知れないであろう．しかし現時点では，このような大規模医療データの分析には，多くの課題が存在する．いかに大規模なデータが蓄積されていても，“Garbage in, Garbage out”という言葉が示すように，データの質が悪ければ，大規模医療データの結果も信用できないものとなってしまう．筆者は，これまで蓄積された膨大な健康医療データを活用する際の課題を克服するために，（1）質の高いデータを収集することによる健康医療データベースの構築，（2）正確なアノテーション情報の蓄積，（3）未観察・未測定交絡への対処，に取り組んできた．本稿では，これらについて概説し，膨大な健康医療データを活用する際の課題を克服する方法を紹介したい．

Ⅰ　質の高い健康医療データベースの構築

筆者は，2020年3月まで国立がん研究センター社会と健康研究センターの在任中，「次世代多目的コホート研究（Japan Public Health Center-based Prospective Study for the Next Generation: JPHC-NEXT）」の横手地域：約3万人地域住民の追跡調査を行ってきた．JPHC-NEXT研究は7県の40-74歳の地域住民11.5万人を対象としたコホート研究であり，生活習慣に関するアンケート調査に協力している[1]．運営には，保健所・自治体・医療機関などに協力依頼し，同意者のうち，5.5万人からは，血液や尿の生体試料の提供を受けており，追跡調査として，死亡を含む異動情報に加え，レセプトデータ，DPCデータ，介護保険データも収集している．サンプルサイズは国民の大部分のデータを有するレセプト情報・特定健診等情報データベース（National Database: NDB）などに比べるとはるかに小さいが，5年ごとの繰り返し調査による社会的経済的な要因・生活習慣や生体試料など詳細な経時的データが得られることが特徴である．このような代表性のあるサンプルにおいて，詳細なデータを収集することにより，選択バイアス・情報バイアス・交絡の影響を最小限にして内的妥当性の高い研究結果を得ることにより，大

後藤　温，横浜市金沢区瀬戸22-2（〒236-0027）横浜市立大学 医学群／大学院　データサイエンス研究科ヘルスデータサイエンス専攻



（文献４を一部改変）

図１：重症低血糖と心血管疾患の関連に関するメタアナリシス

規模医療データから正しい結論を出すことを可能にすることが期待される．今後，横浜市立大学が質の高いコホートデータや病院データを収集し，質の高い健康医療データベースの構築が可能となると考えている．

Ⅱ　正確なアノテーション情報の蓄積

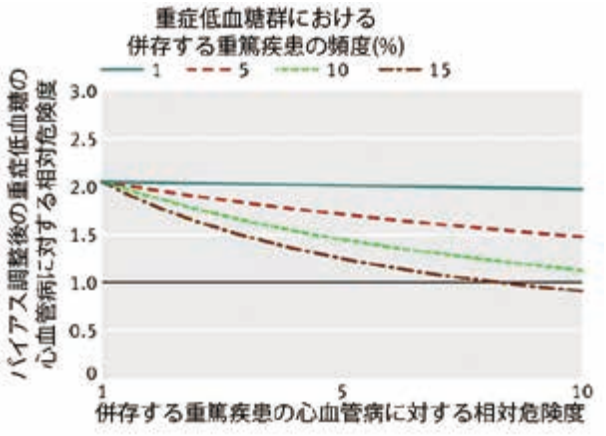
大規模医療データの活用のみならず，「人工知能（Artificial Intelligence：AI）」技術の活用の鍵とされているのが，データの正確なアノテーションである．大規模医療データに含まれるICD-10などの病名コードや薬剤や診療行為コードなどのデータはデータベースに格納しやすいし，分析も容易であるが，情報に漏れや誤りがあることが多い．全国がん登録など一定の基準で網羅的に登録された情報が他のデータベースと突合できるような仕組みを作ることが理想的である．しかし国民からの理解や医療版マイナンバーがないことによる技術的な側面から，現時点ではまだ突合できない状況である．そのような仕組みができるまでは，正確にデータにアノテーションをすることにより，コンピューターに「教師データ」を与えることで，バイアスの小さい推計結果を得ることが期待される．筆者は，これまでのコホートデータのバリデーション研究に取り組んできた経験を活かし[2, 3]，糖尿病やがんをレセプトデータやDPCデータのみで定義するアルゴリズムの開発を行っている．このように医療データベースの弱点を克服し，世界に誇る質の高い医療提供体制の下で，蓄積されてきた膨大な医療データを利活用するための基盤を構築していくことが重要であろう．

Ⅲ　未観察・未測定交絡への対処

観察されたデータを用いて因果推測を行う際の最大の難題は未観察・未測定交絡である．これにより，ある医学的行為が健康アウトカムに影響を与えると誤って結論を出してしまう可能性がある．一般に，ランダム化比較試験を行うことにより，因果関係の評価を行うことが可能であるが，費用的・倫理的観点から実施困難であることが多い．

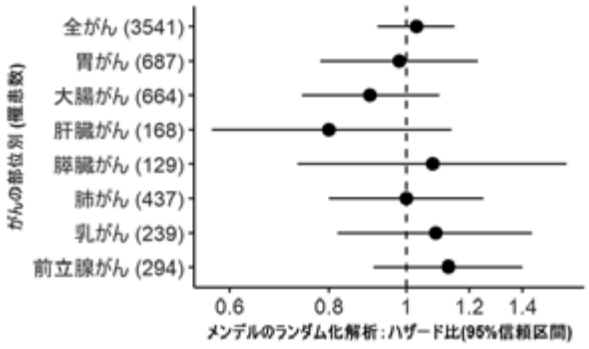
筆者は，観察データにおける因果推測の課題を克服するために，定量的バイアス分析を適用し，糖尿病治療に伴う重症低血糖と心血管疾患リスクとの関連をBMJ誌に報告した[4]．本研究では，「低血糖」「２型糖尿病」「心血管疾患」というキーワードで電子データベース（Embase, MEDLINE, Cochrane library, Web of Science）を用いて網羅的に文献を検索し，3, 443件を抽出した．そのうち，低血糖と心血管疾患リスクの関連をみていない研究，１型糖尿病患者が対象であった研究などを除外したところ，最終的に６件が当該目的に合致し，対象者数は合計903, 510名であった．全６件の研究で重症低血糖は心血管疾患リスク上昇と関連していた．変量効果モデルでメタアナリシスした結果，重症低血糖発生群では非発生群と比べ，心血管疾患発生の相対危険度は2. 05であった（95%信頼区間 1. 74－2. 42）（図１）．

重症低血糖は心血管疾患リスクと正に関連していたが，観察研究のメタアナリシスで得られた結果であるため，バイアスや交絡の存在により系統的に真値からずれた結果を観察している可能性がある．例えば，この関連は併存する重篤疾患による交絡で説明できるかもしれない．



（文献４を一部改変）

図２：重症低血糖と心血管疾患の関連に関するバイアス分析



（文献５を一部改変）

図３：メンデルのランダム化による糖尿病とがんリスクの関連

併存する重篤疾患（重症肝障害，慢性腎不全など）が重症低血糖と心血管疾患を発生させる場合，重篤疾患の存在を考慮した解析（重篤疾患を統計学的に調整すること）をする必要がある．しかし，メタアナリシスに含められた研究では重篤疾患の存在が考慮されていなかったため，重症低血糖と心血管疾患との関連は見かけ上の関連かもしれない．ランダム化比較試験であれば，バイアスや交絡の問題を解決できるが，重症低血糖群と非重症低血糖群に割り付けるランダム化比較試験は倫理的観点から実施困難である．一方，バイアス分析を用いたメタアナリシスは，バイアスを考慮した上で重症低血糖による心血管疾患リスクを定量的に推定することが可能である．そこで，バイアス分析を実施したところ，臨床的に想定できる範囲で重篤疾患の存在を考慮しても重症低血糖は心血管疾患と正に関連する（バイアス調整後の重症低血糖の心血管疾患に対する相対危険度＞１）ことが示された（図２）．重症低血糖と心血管疾患との関連を説明するためには，重症低血糖発生群で重篤疾患併存頻度が10倍以上で，かつ重篤疾患と心血管疾患との相対危険度が10倍以上である必要があった．

この研究により，低血糖を避ける糖尿病治療の重要性が広く認識されるようになり，国内外の論文・診療ガイドライン・教科書で広く引用されている．

さらに筆者らは，糖尿病とがんとの間の複雑な関係を紐解くべく，Mendelian Randomization（MR）法を適用した．MR法は，近年，交絡要因による影響を小さくすることにより観察研究における因果推測を行う方法として注目されている．これは遺伝子型を操作変数として交絡要因による影響を小さくすることにより，曝露因子と疾病との間の因果関係について検討する操作変数法の一種である．

従来の手法で解析したコホート研究では，糖尿病はがんリスク上昇と一貫して関連することが示されていたが，残余交絡や因果の逆転などの可能性があるため，糖尿病ががんの原因であるか否かは明らかでなかった．そこで筆者らは，MR法を適用することにより，糖尿病ががんの真のリスク因子であるかを検討した[5]．コホート研究で血液を提供した40～69歳の男女約３万３千人を，2009年末まで追跡し，がん罹患が確認された3,541人と無作為に選んだ対照グループ10,536人のゲノム網羅的タイピングを実施した．1000万以上の遺伝子多型の中から２型糖尿病と関連する29個の遺伝子多型を抽出し，MR法により糖尿病とがん全体および部位ごとのがんリスクとの関連を分析した．その結果，糖尿病の有病率が倍になることによるがん罹患リスクは，がん全体で1.03倍（95％信頼区間：0.92～1.15），大腸がんで0.90倍（0.74～1.10），肝臓がんで0.80倍（0.57～1.14），膵臓がんで1.08倍（0.73～1.59）と推計された（図３）．さらに日本人おける6,692人の大腸がん症例と27,178人の対照による大規模データを用いて解析したところ，大腸がんのリスクは1.00倍（0.93～1.07）であった．

この研究は，MR法を用いて２型糖尿病と全がんおよび部位別のがんリスクとの関連を検討した世界で初めての研究である．MR法を応用すると有望な創薬ターゲットを同定したり，複数の因子が病態に及ぼすメカニズムを明らかにしたりすることも可能であり，今後，様々研究課題に対して，大規模なサンプルを用いたMR研究を展開することを計画している．

おわりに

一般に，質の高い医学的エビデンスを得ることは非常に難しい．サンプルサイズが大きく，代表性があるデータであっても，詳細な臨床情報，生活習慣や社会経済状況などの情報が得られていないことが多い．さらに，個々の研究課題に対して，選択すべき研究デザインや分析法は一つとは限らない．正しい結論を得るためには正確で詳細なデータを収集することに努めるとともに，さまざまな異なる方法で分析し，互いの欠点を補って，真実に迫っていこうとする「triangulation of evidence」の考え方が重要である．今後，公衆衛生学や臨床医学のさまざまな課題に対して，データの利点と限界を把握した上で，適切なデータの加工と分析を行うとともに，この分野における人材育成にも尽力していきたい．

謝　辞

本研究に際して，ご指導，ご助言を下さった寺内康夫先生，野田光彦先生，津金昌一郎先生，岩崎　基先生，澤田典絵先生，Simin Liu先生，Onyebuchi A. Arah先生をはじめとする先生方に心より感謝申し上げます．

文　献

1）Sawada N, Iwasaki M, Yamaji T, Goto A, Shimazu T, Inoue M, et al.: The Japan Public Health Center-based Prospective Study for the Next Generation（JPHC-NEXT）: Study Design and Participants. J Epidemiol, **30**（1）: 46－54, 2020.

2）Goto A, Morita A, Goto M, Sasaki S, Miyachi M, Aiba N, et al.: Validity of diabetes self-reports in the Saku diabetes study. J Epidemiol, **23**（4）: 295－300, 2013.

3）Goto A, Goto M, Noda M, Tsugane S.: Incidence of type 2 diabetes in Japan: a systematic review and meta-analysis. PLoS One, **8**（9）: e74699, 2013.

4）Goto A, Arah OA, Goto M, Terauchi Y, Noda M: Severe hypoglycaemia and cardiovascular disease: systematic review and meta-analysis with bias analysis. BMJ, **347**: f4533, 2013.

5）Goto A, Yamaji T, Sawada N, Momozawa Y, Kamatani Y, Kubo M, et al.: Diabetes and cancer risk: A Mendelian randomization study. Int J Cancer, **146**（3）: 712－719, 2020.

Abstract

CLINICAL RESEARCH THAT ADDRESSES THE CHALLENGES OF ANALYZING LARGE-SCALE MEDICAL DATA

Atsushi Goto

Department of Health Data Science, Graduate School of Data Science, Yokohama City University

In recent years, omics data such as genomic information and metabolomic information, as well as health screening and medical data, have been accumulated in large-scale databases. In the healthcare field, it is expected that various problems can be solved by using this large-scale medical data. However, there are many challenges in analyzing such large-scale medical data at present. To introduce methods to overcome the challenges of using vast amounts of healthcare data, this article outlines (1) the construction of healthcare databases by collecting high-quality data, (2) the accumulation of validation data, and (3) addressing unobserved and unmeasured confounding.